Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1004 (2003) 13–28
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

New approach based on fuzzy logic and principal component analysis for the classification of two-dimensional maps in health and disease
Application to lymphomas

Emilio Marengo[a,*], Elisa Robotti[a], Pier Giorgio Righetti[b], Francesca Antonucci[b]
[a]*Dipartimento di Scienze e Tecnologie Avanzate, Università del Piemonte Orientale, 15100 Alessandria, Italy*
[b]*Dipartimento Scientifico e Tecnologico, Università di Verona, Strada le Grazie 15, 37134 Verona, Italy*

Abstract

Two-dimensional (2D) electrophoresis is the most wide spread technique for the separation of proteins in biological systems. This technique produces 2D maps of high complexity, which creates difficulties in the comparison of different samples. The method proposed in this paper for the comparison of different 2D maps can be summarised in four steps: (a) digitalisation of the image; (b) fuzzyfication of the digitalised map in order to consider the variability of the two-dimensional electrophoretic separation; (c) decoding by principal component analysis of the previously obtained fuzzy maps, in order to reduce the system dimensionality; (d) classification analysis (linear discriminant analysis), in order to separate the samples contained in the dataset according to the classes present in said dataset. This method was applied to a dataset constituted by eight samples: four belonging to healthy human lymph-nodes and four deriving from non-Hodgkin lymphomas. The amount of fuzzyfication of the original map is governed by the σ parameter. The larger the value, the more fuzzy the resulting transformed map. The effect of the fuzzyfication parameter was investigated, the optimal results being obtained for $\sigma = 1.75$ and 2.25. Principal component analysis and linear discriminant analysis allowed the separation of the two classes of samples without any misclassification.

Keywords: Fuzzy logic; Principal component analysis; Linear discriminant analysis; Proteomics; Chemometrics

1. Introduction

It is commonly acknowledged that changes in the protein content of cells and of biological fluids are often involved in the onset and evolution of particular diseases such as Alzheimer disease [1], cancer [2–5], Creutzfeldt–Jakob disease [6] and leukaemia [7]. The investigation of the proteins synthesised by a particular cell type or contained in a biological fluid becomes thus fundamental for the study of the disease. Every specimen may contain thousands of different proteins: this large amount of components hampers proper separation of all species. This problem has partially been solved by the early development of the two-dimensional electrophoretic separations, certainly the most widely used analytical methods in proteomics [8,9]. This technique allows an efficient separation of the protein content of a particular cell or fluid, producing a two-dimensional (2D) image of the protein species present in the sample under investigation. The separated proteins

*Corresponding author. Fax: +39-0131-287-416.
E-mail address: marengoe@tin.it (E. Marengo).

0021-9673/03/$ – see front matter
doi:10.1016/S0021-9673(03)00852-5

appear as spots spread all over the two-dimensional map (one dimension for the pH elution, the other for the proteins mass separation), dubbed ''2D-PAGE'', where PAGE = polyacrylamide gel electrophoresis. Since a 2D-PAGE may be considered as a snapshot of the state of the cell investigated according to its protein content, it may be used for both diagnostic and prognostic purposes: by investigating the differences occurring between the 2D-PAGE gels of control and pathological individuals, it is possible to classify the patients accordingly or even to capture the evolution of the same disease. The solution of the problem concerning the comparison of maps belonging to different individuals becomes the fundamental key for the application of this powerful technique to diagnostic and prognostic purposes.

The comparison of different 2D maps is a hard problem to solve, because of the complex specimen investigated and of the instrumental technique applied. The problems which characterise the comparison can be summarised in:

(i) a complex specimen that leads to complex maps, often containing thousands of spots with the additional appearance of spurious spots due to side reactions;

(ii) a complex sample pre-treatment, characterised by several purification and extraction steps, that causes an increase in the uncertainty and a subsequent decrement of the reproducibility of the maps;

(iii) the great amount of spots present on the same map, which hampers the identification of the sometimes small differences occurring between affected and healthy individuals;

(iv) the presence of many experimental factors that influence the electrophoretic run (polymerisation conditions, temperature, staining conditions), which lead to differences due not only to real variations between the samples but also to accidental variations due to the experimental steps performed.

These problems can cause a low repeatability of the 2D maps of the same individual: replicated 2D maps may be very different from each other, showing variations in the spot size, shape, position and number.

In the classical approach, the comparison is performed by specific software, e.g. Melanie II or PDQuest [10–12]. In this case each 2D slab gel is analysed by a densitometer which provides, for each 2D map, the level of optical density in each point of the map. The analysis performed by the mentioned software is constituted by the following different steps:

(i) *Spot detection*: the spot detection tool guides the operator through the process of identifying protein spots in the gel image. After a suitable setting of the detection parameters, one can use the same set to detect spots on all gels that were stained, run and scanned under the same conditions.

(ii) *Spot revelation*: the software performs the revelation of the spots independently on each map. Because the image profile of an ideal spot conforms to a Gaussian curve, PDQuest uses Gaussian modelling to create ''ideal'' spots. A Gaussian spot is a precise three-dimensional representation of an original scanned spot to accurately identify and quantify real spots.

(iii) *Matching of the maps*: the maps are matched one to the others so as to reveal the common features (spots present in all the maps) and the different ones (spot detected only on some of the samples under investigation). When the compared maps are replicates of the same sample, this step produces a ''synthetic'' map which summarises the common information and which contains the spots present in all the maps compared.

The key point of this procedure is the matching of the maps, which may produce very different results even by varying the software parameters.

The great amount of data resulting from the application of this procedure to a variety of 2D map sets has made it necessary to couple such methods with multivariate techniques, for example, principal component analysis (PCA) [13–16]. In the past, PCA has been often applied to the study of DNA and RNA fragments of several biological systems [17–20] and to the characterisation of gel-electrophoresis patterns belonging to different classes of samples [21–26]; recently, Kovarova et al. [27] have proposed the use of PCA for the classification of proteomic patterns by coupling 2D gel electrophoresis with PCA, for the characterisation of the anticancer activity of bohemine (a new omoleucine-derived synthetic cyclin-dependent kinase inhibitor).

In our laboratory the problem was tackled from a different point of view, by applying a new approach based on the combined use of fuzzy logic [28] and

principal component analysis (PCA) followed by classification methods [16], for the comparison of two-dimensional electrophoretic gels. Fuzzy logic has been recently applied to the recognition of bacteria, generally coupled with the use of artificial neural networks [29,30]. Classification methods, like linear discriminant analysis (LDA), have instead been applied to image retrieval of pixel data [31–33].

The approach proposed here consists in combining the principles of fuzzy logic with multivariate data mining techniques and classification methods. The central point of our approach is that each spot is allowed to smear [34,35]; spots present on each map are no longer univocally positioned since they are substituted by *fuzzy entities*, being transformed into probability functions, thus implicitly taking into consideration the low reproducibility that affects the experimental method applied. The comparison is then performed on the ''fuzzy'' maps obtained from each original map. The samples can be analysed by chemometric methods, like PCA and classification methods, in order to evaluate the differences characterising the various classes of samples taken into account. In this paper, PCA is performed on a set of samples belonging to normal lymph-nodes and to lymphomas; the identification of the two classes of samples and of the characterising differences is then performed by LDA.

2. Theory

2.1. Principal component analysis

PCA is a multivariate statistical method which allows the representation of the original dataset in a new reference system characterised by new variables called principal components (PCs). Each PC has the property of explaining the maximum possible amount of variance contained in the original dataset. The PCs, which are expressed as linear combinations of the original variables, are orthogonal one to each other and can be used for an effective representation of the system under investigation, with a lower number of variables than in the original case. The co-ordinates of the samples in the new system of variables are called *scores* while the coefficient of the linear combination describing each PC, i.e. the weights of the original variables on each PC, are called *loadings*.

2.2. Linear discriminant analysis

LDA [36,37] is a Bayesian classification method that allows the discrimination of the samples present in a dataset considering its multivariate structure. An object $\mathbf{x}$ is assigned to the class g for which the posterior probability $P(g/\mathbf{x})$ is maximum, assuming a Gaussian multivariate probability distribution:

$$P(g|\mathbf{x}) = \frac{P_g}{(2\pi)^{p/2} \cdot |S_g|^{1/2}} \cdot \exp[-0.5(\mathbf{x} - \mathbf{c}_g) \cdot S_g^{-1}(\mathbf{x} - \mathbf{c}_g)] \quad (1)$$

where P_g is the prior probability, S_g is the covariance matrix of class g that, in the case of LDA, is approximated with the pooled (between the classes) covariance matrix, $\mathbf{c}_g$ is the centroid of class g, p is the number of descriptors. The argument of the exponential function is the Mahalanobis distance between the object $\mathbf{x}$ and the centroid of the class g, which takes into consideration the class covariance structure:

$$(\mathbf{x} - \mathbf{c}_g) \cdot S_g^{-1}(\mathbf{x} - \mathbf{c}_g) \quad (2)$$

From the logarithm of the posterior probability, by eliminating the constant terms, each object is classified in the class g if it is minimum, the so-called *discriminant score*:

$$D(g|\mathbf{x}) = (\mathbf{x} - \mathbf{c}_g) \cdot S_g^{-1}(\mathbf{x} - \mathbf{c}_g) + \ln|S_g| - 2\ln P_g \quad (3)$$

The selection of the variables for the LDA models which discriminate the classes present in the dataset was performed by a stepwise algorithm in forward search ($F_{\text{to-enter}} = 4.0$).

2.3. The applied method

Since a 2D map can be considered as a snapshot of the protein content of the cell under investigation, we applied our approach to a set of 2D maps scanned with a GS-710 densitometer (Bio-Rad Labs., Rich-

mond, CA, USA). The approach consists of the following steps:

(i) *Digitalisation of the 2D-PAGE image*; in this step each image is transformed into a grid of 200×200 pixels containing the values of the optic density scaled from 0 to 1.

(ii) *Map de-fuzzyfication*; in this step the values of the optic density smaller than a fixed threshold are cut off and substituted by null values. Then, the grid is matched to the correspondent image, so that a 0 value is assigned where no signal is present and a 1 value is assigned where a signal is detected. The sensitivity to the destaining protocol is eliminated in this step.

(iii) *Map re-fuzzyfication* [34,35]; in this step the spots present in each map are transformed into probability functions. The low reproducibility of two-dimensional electrophoresis is taken into account here and a new grid of 200×200 cells is produced, containing in each cell a value corresponding to the probability of finding a signal in that position.

(iv) *Principal component analysis*; PCA is performed on the fuzzy maps in order to understand which parts of the 2D-PAGE maps contain the same type of information. Moreover PCA can permit the identification of clusters and groups of samples. Finally it provides a size reduction of the dataset since from thousands of original variables only the relevant PCs can be maintained.

(v) *Classification*; the LDA is performed in order to discriminate the classes of samples involved in the study and to identify the spots responsible of the separation of the original dataset into classes. LDA is applied on the relevant PCs obtained from the previous step.

2.3.1. Digitalisation

The digitalisation was performed by Matlab software, using the Image Processing Toolbox on the images obtained by scanning the maps with a GS-710 Densitometer (Bio-Rad). The image of each map is turned into a grid of 200×200 cells, containing in each cell the value of the optic density in the correspondent position, ranging from 0 to 1. The intensity values on the grey scale lower than 0.40 were cut off and substituted by null values, in order to cancel the contribution to the signal given by the background. The value of 0.4 was chosen arbitrarily on the basis of a visual inspection of the digitalised maps. The value of 0.4 proved to be effective for all the 2D-PAGE maps considered.

2.3.2. De-fuzzyfication

The grid containing the values of the optic density is turned into a grid containing only binary values (0—1): 0 if the optic density is below 0.4, 1 in the other cases.

The value of the optic density is influenced by the destaining protocol: transforming each value of the optic density into a binary value corresponds to performing a ''de-fuzzyfication'' of the map, which allows the elimination of the sensitivity to the destaining protocol. The importance of this step is stressed by Fig. 1, which shows the digitalised map for sample HEA1, together with the de-fuzzyfied and the re-fuzzyfied map, as an example.

2.3.3. Re-fuzzyfication

With the de-fuzzyfication step, the information about the spatial imprecision due to the electrophoretic run is lost: to re-introduce it, a re-fuzzyfication step is necessary. The re-fuzzyfication of the maps is a focal point of our approach since, as already remarked, the size, shape and position of the spots on a map may be quite different for electrophoretic runs performed on the same specimen: it appears thus very dangerous to locate precisely and univocally a spot on a map by the two coordinates x and y. In order to consider this effect, each cell containing a 1 in the digitalised image is substituted with a two-dimensional probability function. The statistical distribution used to this purpose is a two-dimensional gaussian function. The probability of the presence of a signal in cell x_i, y_i when a signal is present in the cell x_k, y_k is calculated by the following function:

$$f(\Delta x_{ik}, \Delta y_{ik}) = \frac{1}{2\pi\sigma_x\sigma_y} \cdot e^{-1/2\cdot[(\Delta x_{ik}/\sigma_x)^2+(\Delta y_{ik}/\sigma_y)^2]} \quad (4)$$

where $\Delta x_{ik} = x_k - x_i$ and $\Delta y_{ik} = y_k - y_i$ are the distances between the cell k containing the spot and the cell i where the probability is computed, along each

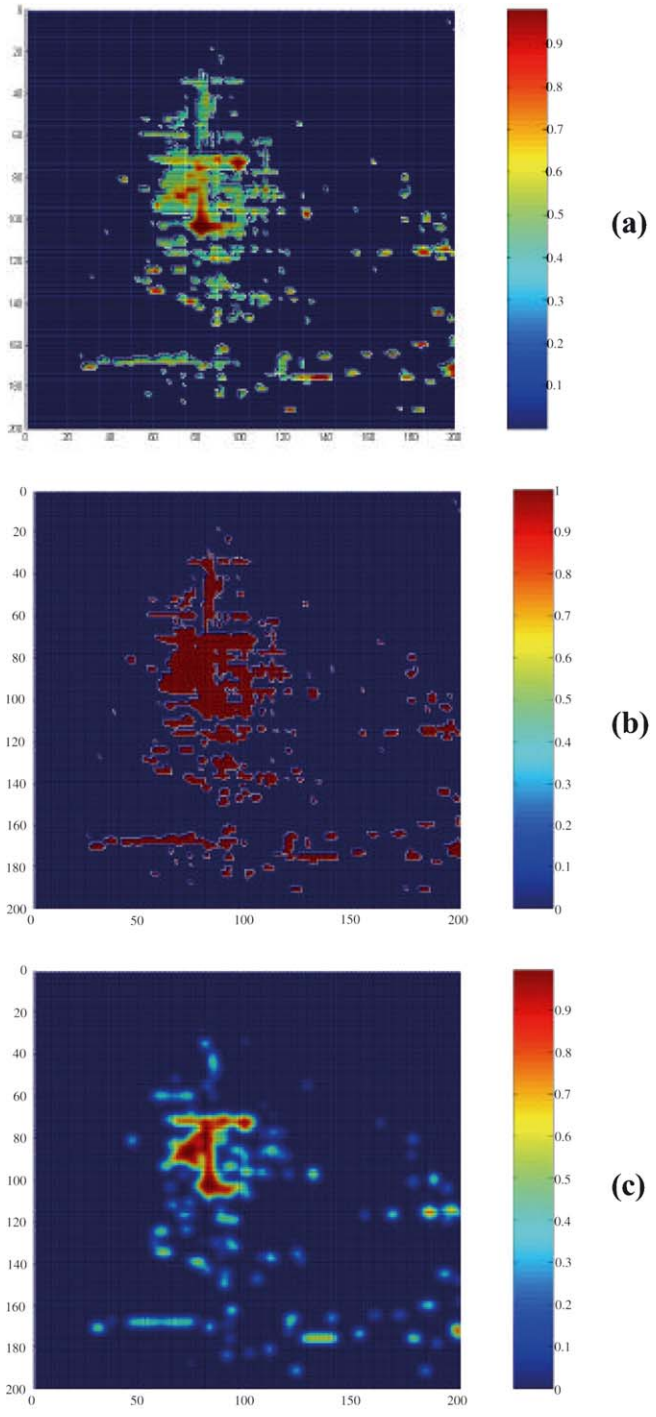

Fig. 1. Digitalised (a), de-fuzzyfied (b) and re-fuzzyfied (c, $\sigma = 1.75$) maps of the control sample HEA1.

axis; σ_x and σ_y are two constants corresponding to the standard deviation of the gaussian function along each of the two dimensions. In the present paper the two parameters σ_x and σ_y are kept identical which corresponds to an identical repeatability of the result with respect to the two electrophoretic runs. So, the parameter which shall be analysed for its effect on the final result is $\sigma = \sigma_x = \sigma_y$. Changing the value of the parameter σ corresponds to modifying the distance at which an occupied cell exerts its effect: high values of σ correspond to a perturbation operating at larger distances and a higher ''fuzzyfication'' level of the maps; a decrease of σ corresponds instead to fuzzy maps more similar to the original 2D-PAGE.

The gaussian distribution was chosen as the best function since the spots can be described as intensity/probability distributions, with the highest intensity/probability value in the centre of the spot itself and decreasing intensities/probabilities as the distance from the centre of the spot increases. Moreover, the integral of the gaussian function on the whole domain of the 2D-PAGE is 1, so that the total signal is blurred but maintained quantitatively coherent.

The value of the signal $\mathbf{S}_i$ in each cell x_i, y_i of the ''fuzzy'' grid is given by the sum of the probability contributions of all neighbouring cells containing signals:

$$S_i = \sum_{j=1,N} f(\Delta x_{ji}, \Delta y_{ji}) \quad (5)$$

The sum runs on all the N cells of the grid, but in dependence on the value of the parameter σ, only the neighbouring cells are affected significantly by the presence of a signal.

Each digitalised image is thus turned into a virtual map containing in each cell the sum of the influence of all the spots of the original 2D-PAGE; these virtual maps can be called *fuzzy matrices* or *fuzzy maps*. The applied transformation corresponds to a gaussian filter applied to each map. The choice to associate the gaussian probability function to each cell instead of to each spot is due to the presence in some maps of large complex spots, whose shape is irregular, so that the substitution of the spot with a gaussian probability function would not describe properly its shape.

3. Materials and methods

The proposed method was applied to a set of 8 real samples, divided in two classes:

(i) 4 samples belonging to healthy human lymphnodes;

(ii) 4 samples belonging to non-Hodgkin lymphomas.

Fig. 2 represents the 8 experimental 2D maps obtained by using the procedure described below.

3.1. Apparatus

The digitalisation of the image was performed by using MATLAB (Mathworks, ver. 6.1); this software was also used for data treatments and for the most part of graphical representations.

Stepwise LDA was performed with STATISTICA (Statsoft, ver. 5.1) and PCA with UNSCRAMBLER (Camo, ver. 7.6). UNSCRAMBLER was also used for the representation of the digitalised and the fuzzy maps and of the scores plots.

3.2. Chemicals and materials

Urea, thiourea, 3-[(cholamidopropyl)dimethylamino]-1-propanesulfonate (CHAPS), iodoacetamide (IAA), tributylphosphine (TBP), and sodium dodecyl sulfate (SDS) were obtained from Fluka (Buchs, Switzerland). Glutaraldehyde, sodium acetate trihydrate and formaldehyde were from Sigma (St. Louis, MO, USA). Ampholines, bromophenol blue and agarose were from Pharmacia-LKB (Uppsala, Sweden). Ethanol, methanol, acetone, acetic acid, silver nitrate and citric acid monohydrate were from Merck (Darmstadt, Germany). Acrylamide, *N*,*N*′-methylenebisacrylamide, ammonium persulfate, *N*,*N*,*N*′,*N*′-tetramethylenediamine (TEMED), Protean isoelectric focusing (IEF) cell, Protean II xi cell, GS-710 densitometer, Mini Trans-Blot electrophoretic transfer cell, and the linear Immobiline dry strips, pH gradient 3–10 (17 cm), were from Bio-Rad Labs. (Hercules, CA, USA).

3.3. Sampling and protein extraction

Sample preparation and solubilization for biopsies

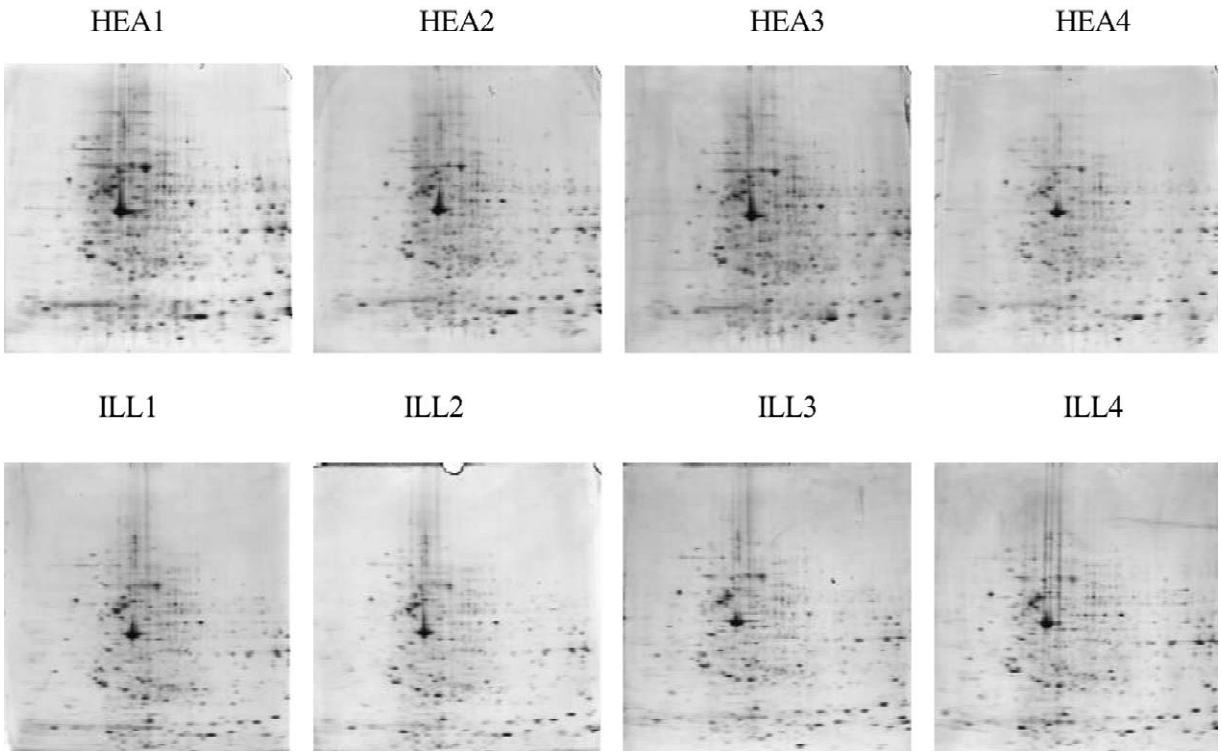

Fig. 2. 2D-PAGE images of the eight investigated samples.

was performed essentially as described by Sanchez et al. [38] for human lymphoma biopsies; the same approach was used also for human healthy lymph-node biopsies (control biopsies). Briefly, ten frozen slices (about 20 μm×5 mm×10 mm) of a human lymphoma biopsy were mixed with 100 μl of 2D solubilising solution containing 7 *M* urea, 2 *M* thiourea, 3% CHAPS, 40 m*M* Tris, 5 m*M* TBP, 1% Ampholines, protease inhibitor, pH ca. 9. After centrifugation, for removal of particulate material [39], 20 m*M* IAA was added to perform complete alkylation of proteins [40,41]. Salts, which can interfere with the 2D separation process and visualization of 2D result, were removed by dialysis. Protein estimation, for each sample, was carried out with the Bio-Rad DC Protein Assay in order to load always the same amount on the immobilized pH gradient (IPG) strips. Samples were stored at −20 °C until used. Samples from three control subjects (1 mg/ml of protein from each sample) were mixed to obtain a representative sample (pool), used to generate four control maps.

3.4. IEF in IPG strips (first dimension)

IEF was performed as follows: seventeen cm long, pH 3–10 IPG strips were rehydrated with protein samples (3 μg/μl for Coomassie-stained gels) mixed with 0.5% Ampholines and a trace of bromophenol blue. Complete sample uptake onto the strips was achieved after 8 h at room temperature. Focusing was carried out at 20 °C for a total of 75 000 V h.

3.5. Separation of proteins in SDS–PAGE (second dimension)

Before the second dimension, IPG strips were equilibrated for 25 min under shaking in a solution containing 6 *M* urea, 2% SDS, 20% glycerol and 0.375 *M* Tris–HCl, pH 8.8. In the second dimension,

home-made vertical acrylamide–bisacrylamide gradient gels (8–18%; dimensions: 182×190×1.5 mm) were used. Electrophoresis was carried out in three subsequent steps: 2 mA/gel for 1 h, 5 mA/gel for 2 h and 10 mA/gel for 10 h, 15 mA/gel till the end of the electrophoresis run. After separation in SDS–PAGE gels, the proteins were visualised by Colloidal Coomassie stain.

4. Results and discussion

Each 2D-PAGE image was automatically digitalised, thus providing a matrix of dimension 200×200 pixels which contains in each cell the value of the staining intensity in the given position (from 0 to 1).

The eight digitalised images obtained were de-fuzzyfied, by turning each matrix containing values ranging from 0 to 1 into new matrices containing only binary values (0—1). The de-fuzzyfied maps were then re-fuzzyfied and the fuzzyfication parameter was increased until the sample's classification proved to become critical. The following values of σ were investigated: 0.25, 0.75, 1.25, 1.75, 2.25, 3.5, 4.75. The first five cases were extensively analysed, while the last two values of sigma, which produced very confused fuzzy maps, were investigated only for searching for the critical value of the parameter for the correct classification of the samples.

Fig. 3 represents, as an example, the fuzzyfication of sample HEA1 for the first five investigated levels of σ. By increasing σ, the influence of each signal on the near cells increases and the whole map becomes more confused.

For comparison, the calculation was also performed on the original matrices, containing the values of the optic density ranging from 0.4 to 1.

4.1. Principal component analysis

The eight fuzzy maps thus calculated can be collected in a matrix of 8×40 000, with the samples on the lines and the cells on the columns: the variables in this case are the probabilities contained in every cell of the fuzzy map which has been unwrapped. This matrix contains a large number of very small values, corresponding to cells far away from any signal. The columns where all values are very small do not contain useful information and can be eliminated without reducing the final results. A study was then performed for the values providing a perfect classification ($\sigma = 1.75$ and $\sigma = 2.25$), to select the proper threshold which allows to reduce the problem dimensionality without affecting the PCA and LDA results. The corresponding critical threshold was the same, namely 0.01, for all the investigated σ values (16 886 variables retained for $\sigma = 1.75$ and 18 309 variables retained for $\sigma = 2.25$). PCA was performed independently for the five investigated values of σ, by applying the critical threshold identified (the results are summarised in Table 1). By increasing the parameter σ, the number of variables maintained in the study increases: this behaviour is due to the fuzzyfication step which improves the number of cells showing a relevant probability of containing a signal.

The percentage of variance explained by the first four PCs increases with the increase of the parameter σ. It is clear that, in all the first five cases considered, the first four PCs are widely sufficient to describe the original dataset, thus greatly reducing the problem dimensionality. These PCs contain, in all five cases, more than 76% of the total variance contained in the original dataset.

Figs. 4 and 5 represent the score plots and the loading plots for the first and fourth principal components as a function of the level of the σ parameter. We focused our attention on the first and the fourth components because these two PCs allowed an exact classification of the samples of the two classes, as shown by the result of LDA, which shall be described later. Fig. 4 represents the score plots of PC_1 and the corresponding loadings for the values of the fuzzyfication parameter corresponding to $\sigma = 0.25$, $\sigma = 0.75$ and $\sigma = 1.25$. In all three cases, the first component allows to distinguish between the two classes of samples: all the samples related to the pathological specimens are grouped at large negative values on PC_1, while the samples related to the healthy individuals, which show a larger variability, are characterised by large positive values on the first component. Sample HEA3, in spite of being on the correct side, is far from the other samples of the same class. Sample HEA4 appears, among the healthy ones, as the most similar to the samples belonging to affected patients. The

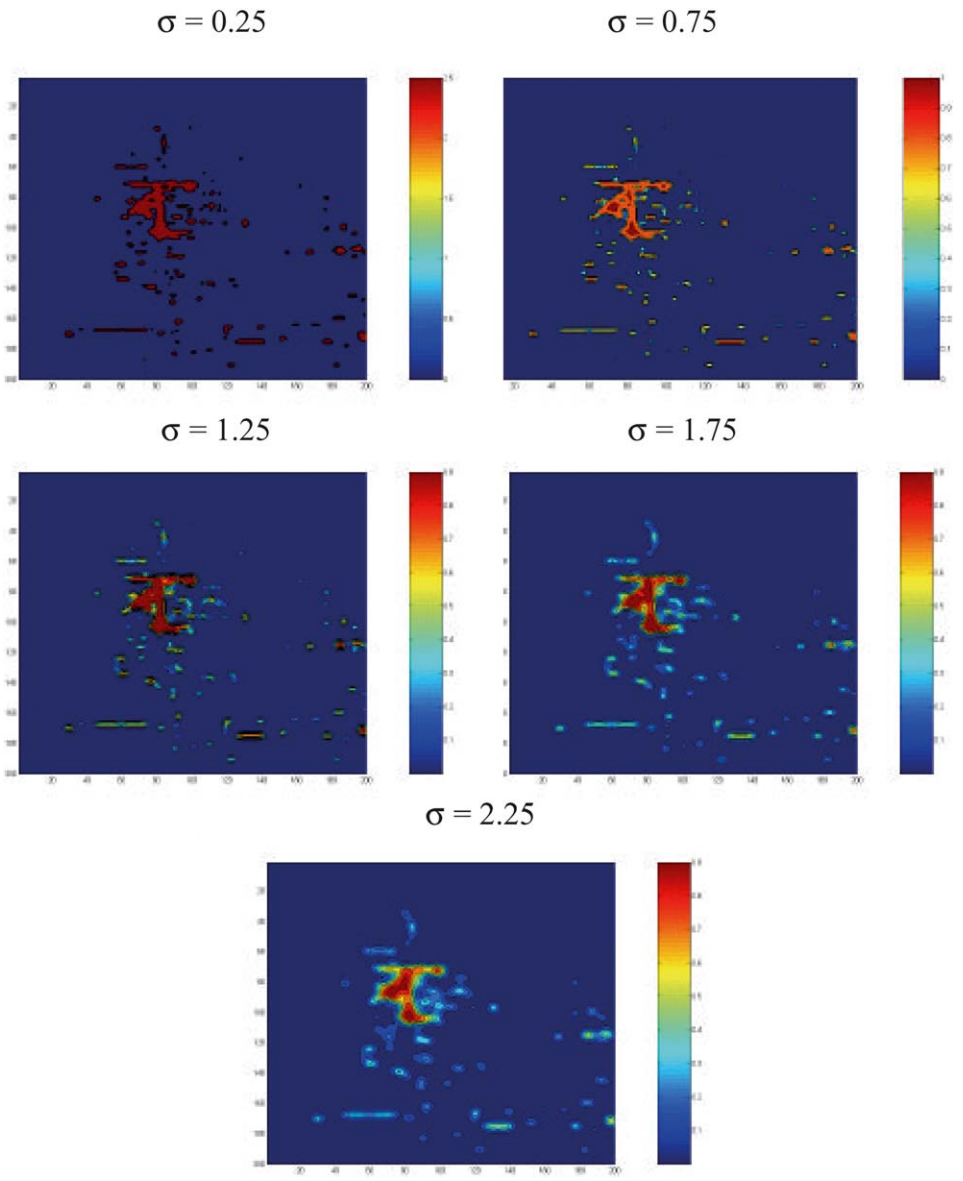

Fig. 3. Fuzzyfication maps of the control sample HEA1 for the five investigated levels of the σ parameter.

loadings are represented in a contour plot isomorphic with the original 2D-PAGE gels. In the map, the colours towards the red indicate large positive loadings while colours towards the blue indicate large negative loadings. The presence of meaningful patterns indicates that the information represented is worth being interpreted. Looking at the loadings plot of the first PC, the pathological samples are characterised by large values of the variables included in the blue zones, whereas the normal samples show large values of the variables indicated with the yellow and the red colours. Thus, the first component allows the discrimination of the two classes of samples (the healthy samples with a larger variability and the affected ones with more repetition) and to identify the variables which mostly characterise the

Table 1
Principal component analysis results for each level of the fuzzyfication parameter σ

σ		% Explained variance	% Cumulative explained variance
–	PC1	30.65	30.65
	PC2	18.22	48.87
	PC3	14.34	63.22
	PC4	13.25	76.47
0.25	PC1	27.24	27.24
	PC2	21.38	48.62
	PC3	14.78	63.41
	PC4	13.39	76.79
0.75	PC1	32.03	32.03
	PC2	21.80	53.83
	PC3	13.74	67.56
	PC4	13.39	80.95
1.25	PC1	37.34	37.34
	PC2	20.94	58.28
	PC3	13.94	72.21
	PC4	11.80	84.01
1.75	PC1	42.75	42.75
	PC2	19.42	62.17
	PC3	14.15	76.33
2.25	PC4	10.12	86.45
	PC1	47.79	47.79
	PC2	17.74	65.52
	PC3	14.16	79.68
	PC4	8.84	88.52

two classes. By increasing the value of the σ parameter, the variability of each class becomes larger, above all for the samples belonging to the normal individuals, but the two classes appear better discriminated.

Fig. 5 represents the loading plots and the score plots for $\sigma = 1.75$ and $\sigma = 2.25$; for these two cases, the loadings for both PC_1 and PC_4 are represented.

According to the score plots of PC_1 versus PC_4, these two components distinguish between the two classes of samples: the pathological subjects at negative values on PC_1 and positive on PC_4 and the healthy samples, at positive values on PC_1 but both positive and negative on PC_4. Again, sample HEA4 appears as the most similar to the pathological ones. The loading plots for PC_1 and PC_4 show, with colours towards blue, the zones which characterise the diseased samples and, with colours towards red, the zones which characterise the control ones.

From the previous analysis, it appears that the two classes of samples are well discriminated from one another by the first and the fourth principal components, for all the values of the investigated σ parameter. In all cases, sample HEA4 seems, among the healthy samples, the most similar to the pathological ones, while sample HEA3 appears as an outlier since it lies far from the other samples of the same class. Increasing the value of σ up to 2.25 causes an increase of the discriminant power of PC_1 and PC_4. Thus, distinguishing between the samples belonging to the two classes investigated becomes easier by increasing the σ parameter.

The variability of the class of the healthy subjects appears larger than that of the ill ones, for all the investigated σ levels.

Since an increase of the parameter seems to improve the separation between the two classes of samples, it becomes fundamental to identify the

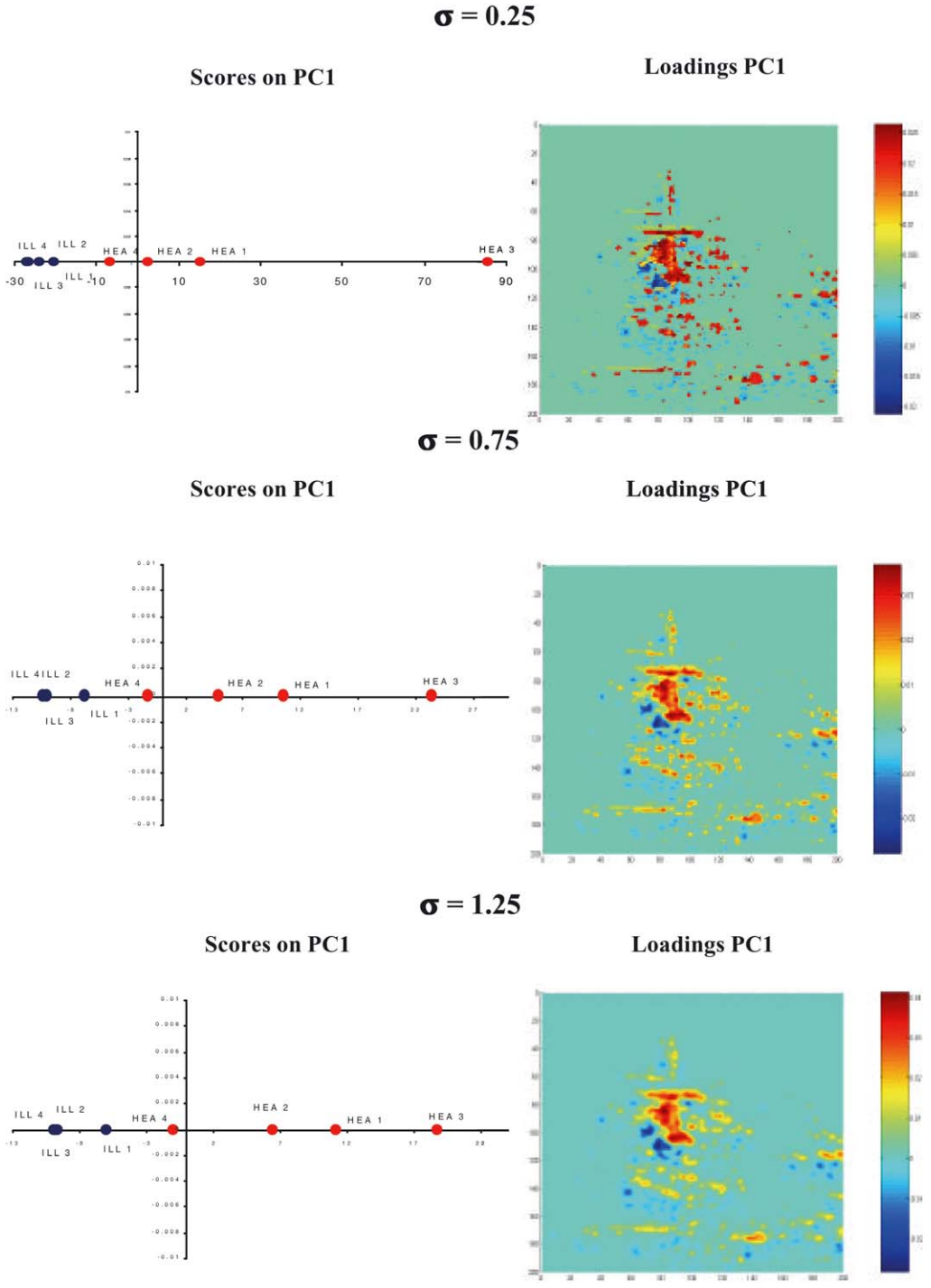

Fig. 4. Score plots and loading plots of PC_1 corresponding to $\sigma = 0.25$, $\sigma = 0.75$ and $\sigma = 1.25$.

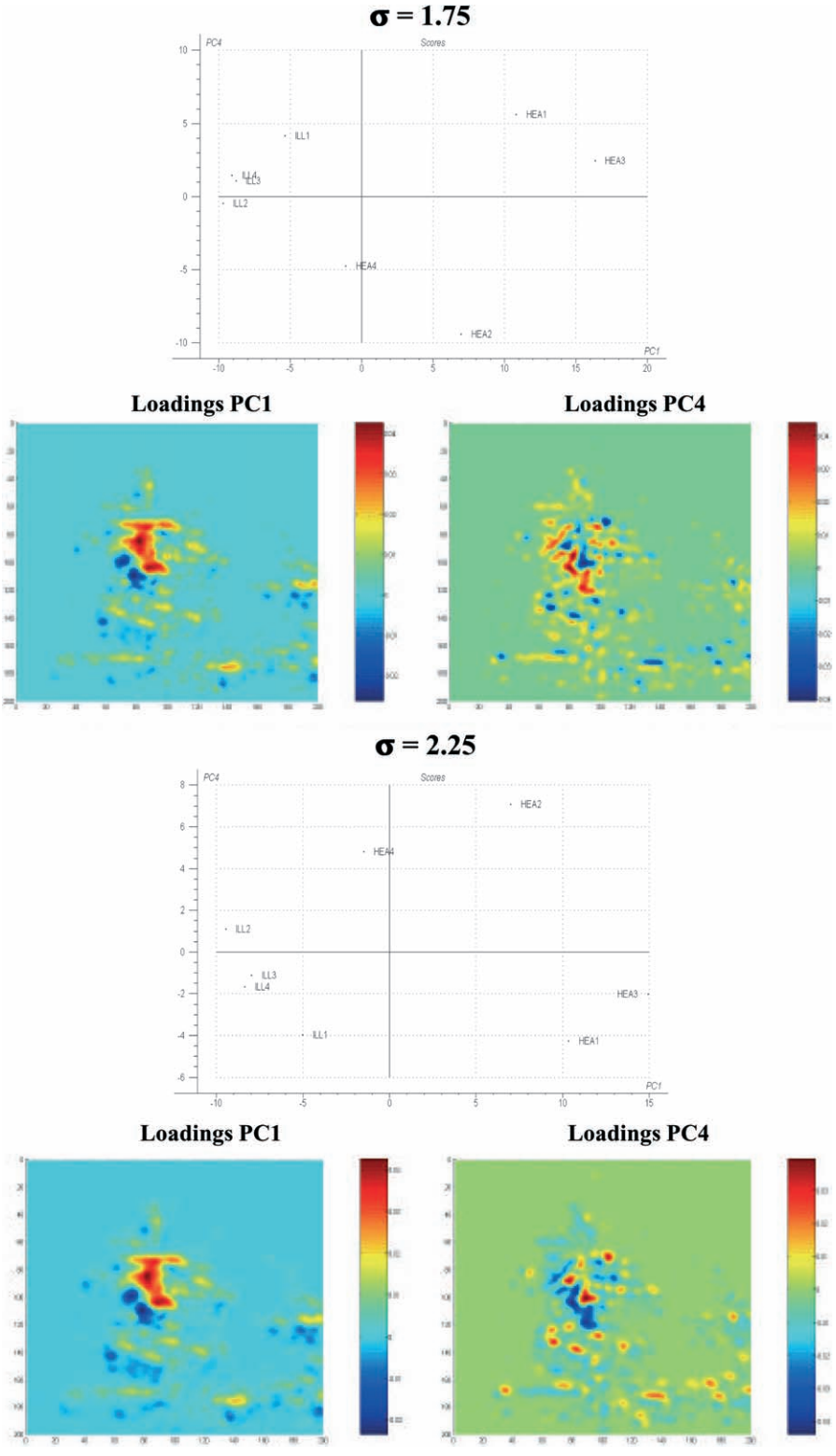

Fig. 5. Score plots and loading plots of PC_1 and PC_4 corresponding to $\sigma = 1.75$ and $\sigma = 2.25$.

range of the σ values which retains a successful separation. The calculation was then performed for larger values of σ, until an inversion of the trend was detected: this effect was recorded for $\sigma = 3.50$ (which required three PCs in the LDA model) and $\sigma = 4.75$ (with one misclassification).

For comparison, it is remarkable that, by performing PCA on the original maps, i.e. on the digitalised matrices containing the values of the optic density ranging from 0.4 to 1 (Table 2), it is not possible to obtain a perfect classification of the samples.

4.2. Linear discriminant analysis

LDA was performed on the dataset of 8 samples, described in terms of the first 5 PCs. The most discriminant variables were selected by a stepwise algorithm in forward search ($F_{\text{to-enter}} = 4.0$). LDA, as a function of the σ parameter, gave the results reported in Table 2. The lowest three values of the fuzzyfication parameter, together with the case where de-fuzzyfication and re-fuzzyfication were not applied, allowed to obtain a NER (non-error-rate) of 87.5% in the classification step (1 misclassification), while for the highest two values of σ, the percentage of well classified samples was 100%. The LDA models include both PC_1 and PC_2 when no fuzzyfication is applied, while they contain only the first PC for $\sigma = 0.25$, $\sigma = 0.75$ and $\sigma = 1.25$. Thus, the classification power increases as the model complexity increases. In all the first three cases, the misclassified sample was HEA4, which was the one nearest to the group of the ill samples in all the score plots considered.

Table 2
Linear discriminant analysis results for each level of the fuzzyfication parameter σ

σ	NER (%)	Wrong classifications
–	87.5	HEA4
0.25	87.5	HEA4
0.75	87.5	HEA4
1.25	87.5	HEA4
1.75	100	–
2.25	100	–
3.50[a]	100	–
4.75[a]	87.5	HEA

[a] Three PCs in the LDA model.

LDA provides a discriminant direction along which the two classes can be distinguished; for the first three levels of the fuzzyfication parameter, this direction is given by the first principal component: samples at positive values on PC_1 belong to the healthy individuals whereas those showing negative values belong to the diseased patients. For the two largest values of the σ parameter, LDA pointed out that both PC_1 and PC_4 are necessary in order to discriminate the two classes of samples; in this case, the discriminant direction is a linear combination of the first and the fourth components.

For all the discriminant models the classification capability was assessed by testing the predictive power with the leave-one-out cross-validation method [42], in fact the available samples are not sufficient for applying a more severe splitting of the dataset into training set and test set. LDA provided 100% of correct predictions up to $\sigma = 3.50$: this value can be taken as the upper limit of applicability of the fuzzyfication step.

4.3. Analysis of the 2D map differences

For the largest value of the fuzzyfication parameter ($\sigma = 2.25$), for which we obtained a perfect classification, we calculated the average fuzzy map of each class, i.e. two ‘‘virtual’’ samples which represent the most typical fuzzy samples of each class. The two maps obtained are graphically represented in Fig. 6. Fig. 7 shows their difference. It is easy to identify the zones of the 2D maps responsible for the differences occurring between healthy and diseased subjects. Fig. 7 represents, at negative values, the spots which characterise the class of affected individuals and, at positive values, the spots which characterise the healthy ones. The healthy samples appear richer in spots and the spots appear also more intense.

5. Concluding remarks

A new method has been developed for the statistical analysis of sets of 2D maps, based on fuzzy logic, principal components analysis and linear discriminant analysis. The method proposed has been applied to a dataset constituted by human healthy

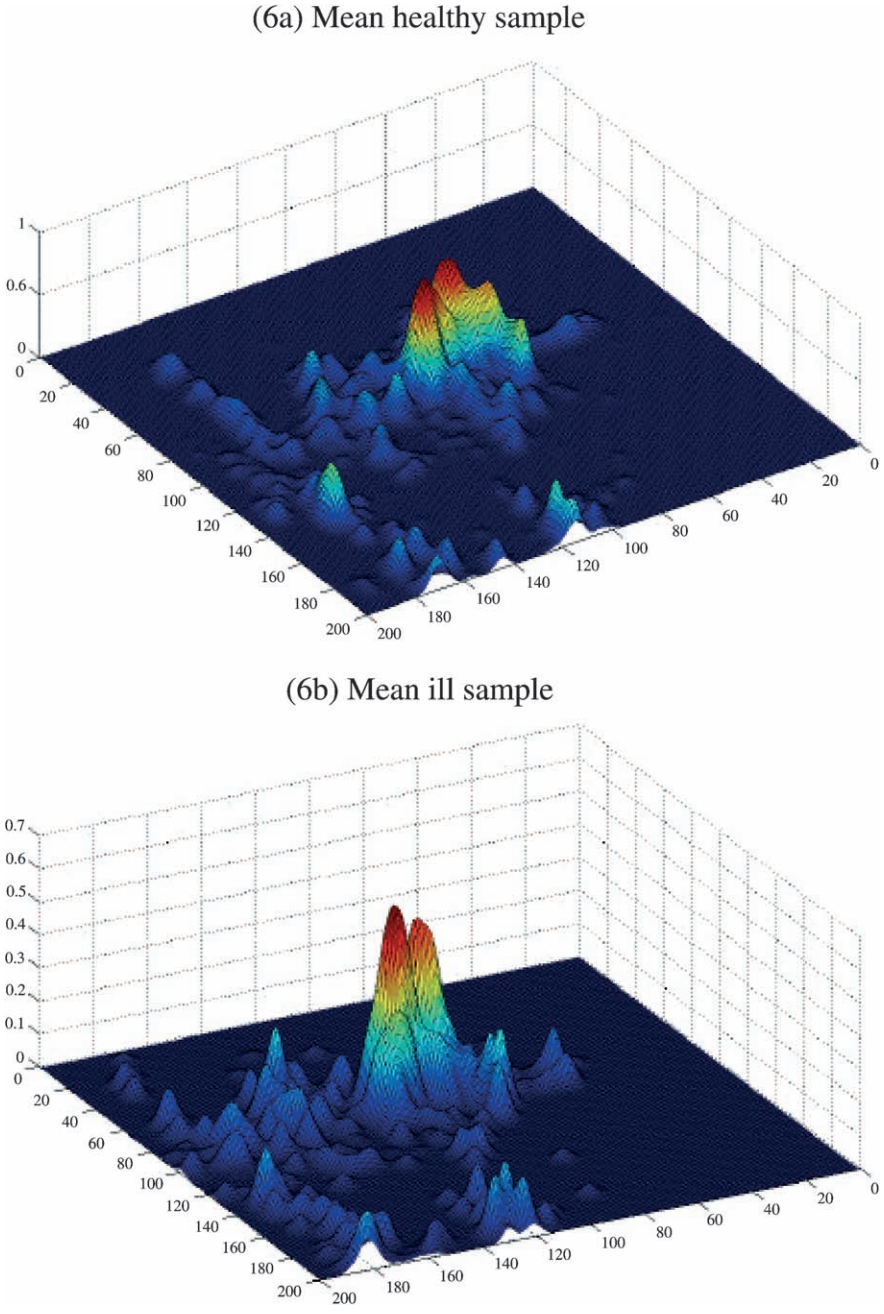

Fig. 6. Mean fuzzy maps of the two classes: healthy (a) and pathological (b) samples.

lymph-nodes and lymph-nodes affected by a non-Hodgkin lymphoma. PCA performed on the dataset allowed the identification of the regions responsible for the differences between the samples and it permitted to highlight not only the differences between the two classes but also the variations occur-

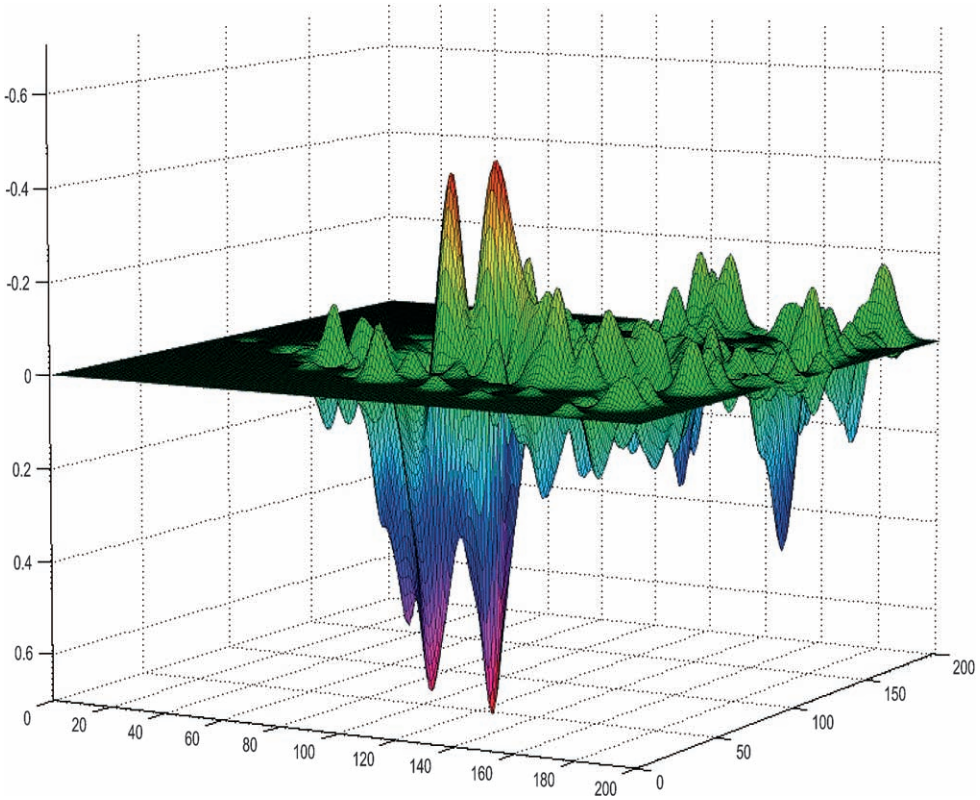

Fig. 7. Fuzzy map obtained as difference of the mean healthy and pathological fuzzy maps.

ring among the samples belonging to the same class. LDA provided the correct classification of the samples by mean of only two PCs (PC_1 and PC_4), thus greatly simplifying the system dimensionality. The application of the method has proved to be better for $\sigma = 1.75$ and $\sigma = 2.25$, showing that the fuzzyfication step is the focal point of the method proposed. The fuzzyfication proved to be successful up to a value of $\sigma = 3.50$, beyond which the confusion of the maps, caused by a too large fuzzyfication level, led to the appearance of misclassification errors. Although the present data are preliminary, they have shown a general agreement with the results obtained by PDQuest analysis of the same set of 2D maps. Moreover, the proposed method, by means of the introduction of the de- and re-fuzzyfication steps, allows to deal with some of the problems, described in the Introduction, which affect two-dimensional gel electrophoresis, namely the complexity of the maps and their low reproducibility. These problems cannot be avoided since they are related to the experimental technique. Work is in progress to refine the model here presented.

Acknowledgements

The authors gratefully acknowledge financial support from MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca, Rome, Italy; COFIN 2000).

References

[1] A. Merched, J.M. Serot, S. Visvikis, D. Aguillon, G. Faure, G. Siest, FEBS Lett. 425 (1998) 225.

[2] L.C. Lawry, J.E. Fothergill, G.I. Murray, Lancet Oncol. 2 (2001) 270.
[3] Y. Katagata, T. Aoki, Y. Hozumi, T. Yoshida, S. Kondo, J. Dermatol. Sci. 13 (1996) 219.
[4] R.J. Simpson, D.S. Dorow, Trends Biotech. 19 (2001) S40.
[5] A. Ardekani, E.H. Herman, F.D. Sistare, L.A. Liotta, E.F. Petricoin, Current Therap. Res. 62 (2001) 803.
[6] D.F. Hochstrasser, S. Frutiger, M.R. Wilkins, G. Hughes, J.C. Sanchez, FEBS Lett. 416 (1997) 161.
[7] J.D. Tissot, F. Spertini, J. Chromatogr. A 698 (1995) 225.
[8] M.R. Wilkins, K.L. Williams, R.D. Appel, D.F. Hochstrasser, Proteome Research: New Frontiers in Functional Genomics, Springer, Berlin, 1997.
[9] P.G. Righetti, A. Stoyanov, M. Zhukov, The Proteome Revisited: Theory and Practice of the Relevant Electrophoretic Steps, Elsevier, Amsterdam, 2001.
[10] F. Hoffmann, K. Kriegel, C. Wenk, Discrete Appl. Math. 93 (1999) 75.
[11] K.H. Lee, Trends Biotech. 19 (2001) 217.
[12] M. Vihinen, Biomol. Engineer. 18 (2001) 241.
[13] J.C. Davis, Statistics and Data Analysis in Geology, Wiley, New York, 1986.
[14] R.G. Brereton, Chemometrics: Application of Mathematics and Statistics to Laboratory Systems, Ellis Norwood, New York, 1990.
[15] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters, Wiley, New York, 1978.
[16] D.L. Massart, B.G.M. Vanderginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, 1988.
[17] R.B. Leonard, J. Mayer, M. Sasser, M.L. Woods, B.R. Mooney, B.G. Brinton, P.L. Newcombgayman, K.C. Carroll, J. Clin. Microbiol. 33 (10) (1995) 2723.
[18] M.L. Johansson, M. Quednau, S. Ahrne, G. Molin, Int. J. System. Bacteriol. 45 (1995) 670.
[19] M.M.B. Couto, J.T.W.E. Vogels, H. Hofstra, J.H.J. Husiintveld, J.M.B.M. Vandervossen, J. Appl. Bacteriol. 79 (1995) 525.
[20] N. Boon, W. De Windt, W. Verstraete, E.M. Top, FEMS Microbiol. Ecol. 39 (2002) 101.
[21] J.E. Alika, M.E. AkenOva, C.A. Fatokun, Genetic Res. Crop Evolution 42 (1995) 393.
[22] N.L. Anderson, R. EsquerBlasco, F. Richardson, P. Foxworthy, P. Eacho, Toxicol. Appl. Pharmacol. 137 (1996) 75.
[23] J. Vohradsky, Electrophoresis 18 (1997) 2749.
[24] H. Kovarova, D. Radzioch, M. Hajduch, M. Sirova, V. Blaha, A. Macela, J. Stulik, L. Hernychova, Electrophoresis 19 (1998) 1325.
[25] S. Lacroix-Desmazes, J. Bayry, N. Misra, M.P. Horn, S. Villard, A. Pashov, N. Stieltjes, R. d'Oiron, J. Saint-Remy, J. Hoebeke, M.D. Kazatchkine, S.V. Kaveri, New Engl. J. Med. 34 (2002) 662.
[26] K. Dewettinck, S. Dierckx, P. Eichwalder, A. Huyghebaert, Lait 77 (1997) 77.
[27] H. Kovarova, M. Hajduch, G. Korinkova, P. Halada, S. Krupickova, A. Gouldsworthy, N. Zhelev, M. Strnad, Electrophoresis 21 (2000) 3757.
[28] M. Otto, Anal. Chim. Acta 283 (1993) 500.
[29] D.Y. Wang, J.M. Keller, C.A. Carson, K.K. McAdoo-Edwards, C.W. Bailey, IEEE Trans. Systems Man Cybern. Part B-Cybernetics 28 (1998) 583.
[30] D.Y. Wang, J.M. Keller, C.A. Carson, Pattern Anal. Appl. 4 (2001) 244.
[31] K. Jensen, C. Kesmir, I. Sondergaard, Electrophoresis 17 (1996) 694.
[32] F.H. Grus, A.J. Augustin, Ophthalmologie 97 (2000) 54.
[33] D. Swets, J. Weng, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 831.
[34] E. Marengo, E. Robotti, V. Gianotti, P.G. Righetti, Ann. Chim. (Rome) 93 (2003) 105.
[35] E. Marengo, E. Robotti, V. Gianotti, P.G. Righetti, D. Cecconi, E. Domenici, Electrophoresis 24 (2003) 225.
[36] R.A. Eisenbeis, Discriminant Analysis and Classification Procedures: Theory and Applications, Lexington, 1972.
[37] W.R. Klecka, Discriminant Analysis, Sage Publications, Beverly Hills, 1980.
[38] J.C. Sanchez, R.D. Appel, O. Golaz, C. Pasquali, F. Ravier, A. Bairoch, D.F. Hochstrasser, Electrophoresis 16 (1995) 1131.
[39] T. Rabilloud, Electrophoresis 17 (1996) 813.
[40] B. Herbert, M. Galvani, M. Hamdan, E. Olivieri, J. McCarthy, S. Pedersen, P.G. Righetti, Electrophoresis 22 (2001) 2046.
[41] M. Galvani, M. Hamdan, B. Herbert, P.G. Righetti, Electrophoresis 22 (2001) 2058.
[42] S. Wold, Technometrics 20 (1978) 397.